CONSIDERING MODEL UNCERTAINTY IN RANDOM REGRESSION MODELS BY MEANS OF BAYESIAN VARIABLE SELECTION

J.P. Steibel[1] and F. Grignola[2]

[1] Facultad de Agronomía, Universidad de Buenos Aires, Cap. Fed., Argentina
[2] Monsanto Company, Saint Louis, MO, USA

INTRODUCTION

In animal breeding, many traits admit repeated measurements or test-day records over time. In order to model the expectation and covariance of these characters as a function of time, the interest in using test-day models (TDM) in the context of the Gaussian mixed linear model has increased in recent years. Even though the advantages of TDM are well known, a problem that frequently arises is the choice of suitable number terms in the linear function. Jensen (2001) discussed various strategies about model choice and, within the Bayesian framework, he suggested Bayesian model averaging to consider the uncertainty around the model for the prediction of breeding values. In this work we implement a simple and flexible Bayesian approach proposed by Kuo and Mallick (1998) to subset a pre-specified set of covariates that best describe the trait of interest in a random coefficient regression model (RRM). The posterior probability of each regressor entering the model is computed using the Gibbs sampling algorithm. The method is illustrated with a simple example where the variable selection strategy is limited to the fixed effects.

METHODS

The method of Kuo and Mallick (1998) expands the usual regression model to include an indicator variable for each predictor considered. In standard notation and for a RRM,

$$y_{jkl} = \sum_{m=0}^{nb} b_{km}\gamma_{km} z_{jlm} + \sum_{m=0}^{nu} u_{jm}\delta_{km} z_{jlm} + \sum_{m=0}^{nu} p_{jm}\theta_{km} z_{jlm} + e_{jkl}$$

where y_{jkl} is the record l for animal j in contemporary group k, z_{jlm} is the known covariate m for animal j and record l associated with time, b_{km} are the "fixed" regression coefficients that describe the average curve for contemporary group (CG) k, u_{jm} and p_{jm} the additive genetic and permanent environmental random regression coefficients for animal j, and γ, δ and θ are indicator variables (I). In matrix notation

$$\mathbf{Y} = \mathbf{X}\Gamma\mathbf{b} + \mathbf{Z}\Delta\mathbf{u} + \mathbf{W}\Theta\mathbf{p}, \text{ with } \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{u} \\ \boldsymbol{p} \\ \boldsymbol{e} \end{pmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{b_0} \\ \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}; \begin{bmatrix} \boldsymbol{D_0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G}\otimes\boldsymbol{A} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{P}\otimes\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{R} \end{bmatrix} \right)$$

where $\mathbf{D}_0$, $\mathbf{G}$ and $\mathbf{P}$ are the covariance matrices for the regressors describing the "fixed", additive genetic and permanent environmental effects, $\mathbf{A}$ = additive relationship matrix, Γ= Diag(γ_{km}), Δ= Diag(δ_{jm}), Θ= Diag(θ_{jm}) and $R = I\sigma^2_e$.

Prior distributions are : $\mathbf{G}$ ~Inv-Wishart$_{nu}((\nu_a, (\nu_a \mathbf{S_a})^{-1})$; $\mathbf{P}$ ~Inv-Wishart$_{nu}((\nu_p, (\nu_p \mathbf{S_p})^{-1})$; σ^2_e ~I-Gamma(ν_e, s_e); γ, δ and θ are i.i.d. with Binomial distribution Bi(1,p_i), where p_i = prior probability of including the i^{th} regression coefficient in the model. The joint posterior density is $f(\Gamma,\beta,\Delta,\Theta p,\sigma^2_a,\sigma^2_p,\sigma^2_e|\mathbf{y}) \propto f(\mathbf{y}|\Gamma,\beta,\sigma^2_e)\, f(\Gamma)\, f(\beta)\, f(\sigma^2_e)\, f(\Delta) f(\mathbf{u}|\mathbf{G}) f(\mathbf{G})\, f(\Theta) f(\mathbf{p}|\mathbf{P}) f(\mathbf{P})$, and the posterior conditional distributions are

$$\begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{u} \\ \boldsymbol{p} \end{pmatrix} \Big| \boldsymbol{Y, G, P, R}, \Gamma, \Delta, \Theta \sim N\left(\boldsymbol{C}^{-1}\boldsymbol{r}; \boldsymbol{C}^{-1}\right), \boldsymbol{r} = \begin{pmatrix} \Gamma' \boldsymbol{X' R^{-1} Y} + \boldsymbol{D_0^{-1} b_0} \\ \Delta' \boldsymbol{Z' R^{-1} Y} \\ \Theta' \boldsymbol{W' R^{-1} Y} \end{pmatrix}$$

$$\boldsymbol{C} = \begin{pmatrix} \Gamma' \boldsymbol{X' R^{-1} X}\Gamma + \boldsymbol{D_0^{-1}} & \Gamma' \boldsymbol{X' R^{-1} Z}\Delta & \Gamma' \boldsymbol{X' R^{-1} W}\Theta \\ \Delta' \boldsymbol{Z' R^{-1} X}\Gamma & \Delta' \boldsymbol{Z' R^{-1} Z}\Delta + \boldsymbol{G^{-1}} \otimes \boldsymbol{A^{-1}} & \Delta' \boldsymbol{Z' R^{-1} W}\Theta \\ \Theta' \boldsymbol{W' R^{-1} X}\Gamma & \Theta' \boldsymbol{W' R^{-1} Z}\Delta & \Theta' \boldsymbol{W' R^{-1} W}\Theta + \boldsymbol{G^{-1}} \otimes \boldsymbol{A^{-1}} \end{pmatrix}$$

$$\left[\boldsymbol{G}\middle|\boldsymbol{Y, b, u, p, P, R}, \Gamma, \Lambda., \Theta\right] = IW\left(q_g + \nu_a; \left(\boldsymbol{u' A^{-1} u} + \nu_a \boldsymbol{S}_a\right)^{-1}\right)$$

$$\left[\boldsymbol{P}\middle|\boldsymbol{Y, b, u, p, G, R}, \Gamma, \Lambda., \Theta\right] = IW\left(q_p + \nu_p; \left(\boldsymbol{p' p} + \nu_p \boldsymbol{S}_p\right)^{-1}\right)$$

$$\left[\sigma^2_e\middle|\boldsymbol{Y, b, u, p, G, P, R}, \Gamma, \Lambda., \Theta\right] = IG\left(n + \nu_e; \boldsymbol{\hat{e}' \hat{e}} + \nu_e \cdot \boldsymbol{S}_e\right), \text{ with } \boldsymbol{\hat{e}} = \boldsymbol{Y} - \boldsymbol{X}\Delta\boldsymbol{b} - \boldsymbol{Z}\Delta\boldsymbol{u} - \boldsymbol{W}\Theta\boldsymbol{p}\,.$$

The posterior distribution of each indicator variable (I_i) is : (I_i |***Y, b, u, p, G, P, R, I*$_{-i}$**)~Bi(1,ω_i)

$$\varpi_i = \frac{p_i \cdot \exp(\boldsymbol{e^{*'} R^{-1} e^{*}})}{(1-p_i)\cdot \exp(\boldsymbol{e^{**'} R^{-1} e^{**}}) + p_i \cdot \exp(\boldsymbol{e^{*'} R^{-1} e^{*}})}\,,$$ e^* = residual if I_i=1 and e^{**} = residual if I_i=0

The method allows to select a different linear function for each effect in the model or to force the same function in some of them.

Linear growth model example. We present a simplified example where the uncertainty around the model is limited to the "fixed" effects. The data came from the beef cattle selection experiment included in the DFREML package (Meyer, 1998). The model was defined as

$$y_{jkl} = \sum_{m=0}^{5}\left(\prod_{i=0}^{m}(\gamma_{ki})b_{km}z_{jlm}\right) + \sum_{m=0}^{5} u_{jm}z_{jlm} + \sum_{m=0}^{5} p_{jm}z_{jlm} + e_{jkl}$$

Note that in this model the specification of the indicator variables is such that it is only possible to select complete polynomials. If one I_i = 0, regressors of higher order will be excluded. Legendre polynomials were used to model the "fixed" and random curves, where *m* determines the number of polynomial terms in the model. The contemporary groups represented 10 paddocks and the maximum *m* allowed in the linear function was 5. For the additive genetic and permanent environmental effects *m* = 2.

Priors. For variance components we considered flat priors : p(**G**), p(**P**), $p(\sigma^2_e) \propto$ constant. However, for **b** we considered alternative priors because previous experience with fixed effects models indicated that the posterior probability of the indicator variables is sensitive to the choice of the priors for **b** (Steibel and Grignola, 2000 ; 2001). These priors were the mean and variance of each regressor obtained from a previous run considering the full model. The resulting variance of each regressor was multiplied by a constant, k = 1, 10, 100, n_i (n_i :

number of observations corresponding to each regressor) to vary the amount of information in the prior.

Gibbs sampling. A blocked Gibbs sampling algorithm was used to estimate the linear parameters, and indicator variables were sampled one at a time. The algorithms of Raftery and Lewis (1992) were used to assess convergence. In order to estimate the quantile 0.5 of each parameter with a precision of 0.01, a burn-in period of 500 iterations and a sample size of 50000 iterations were required.

RESULTS AND DISCUSSION

Posterior probabilities. The posterior probabilities of the selected models were affected by the choice of the prior for **b** (figure 1). Based on the variance components estimates and previous experience with this method (Steibel and Grignola, 2001), only results from the analyses using k = 10 and k = n_i for the priors of **b** are presented.

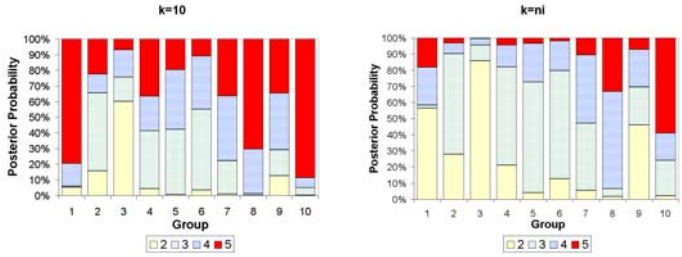

Figure 1. Posterior probabilities for each possible model using different priors for b. Different colors represent the order of the selected polynomial.

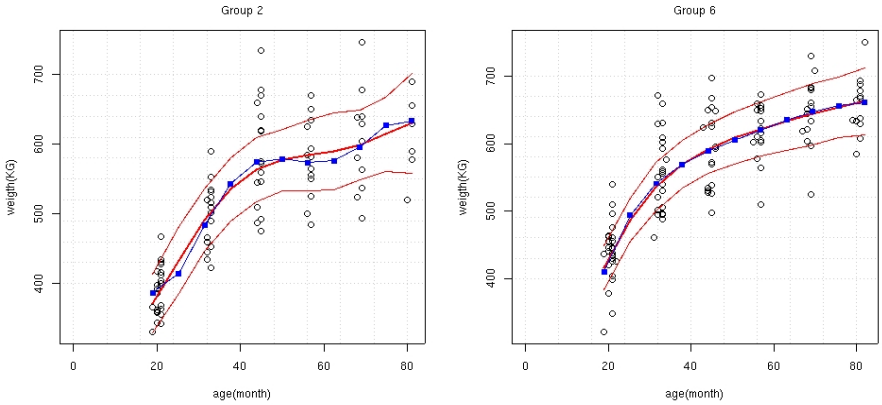

Figure 2. Growth curves posterior means (thick lines) ± 3 STD (thin lines) obtained across models and k = 10, and posterior means for the full model (blue lines with squares)

Average growth curves. Figure 2 shows the posterior probabilities for growth curves averaged across models, for 2 contemporary groups and prior k = 10, and for the complete

model. The results indicate that the curve from the average model performs as good as the full model.

Genetic parameters. Quantiles for the posterior distribution of the variance components presented in table 1 show that the Bayesian selection method gives similar results to those obtained under the full model.

Table 1. Quantiles for the marginal posterior distribution of the variance components corresponding to the full models and models with variable selection

	Full-model						Selection k=10					
	G			P			G			P		
Element	$Q_{0.05}$	$Q_{0.50}$	$Q_{0.95}$	$Q_{0.05}$	$Q_{0.50}$	$Q_{0.95}$	$Q_{0.05}$	$Q_{0.50}$	$Q_{0.95}$	$Q_{0.05}$	$Q_{0.50}$	$Q_{0.95}$
11	2145	3392	4609	482	1216	2279	2144	3371	4562	472	1191	2236
12	265	547	837	-17	158	400	263	541	826	-13	157	392
13	-764	-501	-271	-416	-194	-32	-751	-492	-265	-414	-196	-35
22	68	147	245	24	73	159	65	142	237	23	70	153
23	-119	-59	-1	-52	-1	45	-117	-60	-3	-53	-4	41
33	58	116	200	24	69	140	55	111	193	23	69	138
σ^2_e:	1262	1359	1465	1262	1359	1465	1284	1383	1493	1284	1383	1493

CONCLUSION

The Bayesian variable selection algorithm presented here incorporates the uncertainty about the specification of the linear functions in the inference of the model parameters. It can be considered as an alternative to the Bayesian model averaging proposed by Raftery *et al.* (1997) when the marginal posterior distribution of the parameters are averaged across models. Next steps will include the implementation of the method for the additive genetic and permanent environmental effects.

REFERENCES

Jensen, J. (2001) *J. Dairy Sci*. 84 : 2803-2812.

Kuo, L. and Mallick, B. (1998) *Sankhya*, B. **60** : 65-81.

Meyer, K. (1998) "DFREML User notes".

Raftery, A., Madigan, D. and Hoeting, J. (1997) *JASA* **92** : 179-191.

Raftery, A. and Lewis, S.M. (1992) In "Bayesian Statistics 4". Editors Bernardo J.M. Smith A.F.M. Dawid, A.P. and Berger, J.O. Oxford University Press.

Steibel, J. and Grignola, F. (2000) XVI Reunión Latinoamericana de Producción Animal, III Congreso Uruguayo de Producción Animal. Marzo 28-31, Montevideo, Uruguay.

Steibel, J.P. and Grignola, F. (2001) *Proc. Reunión Científica Grupo Argentino de Biometría*. p14. Argentina.